

Preparing and sharing data with the ELRC-SHARE repository and what happens next

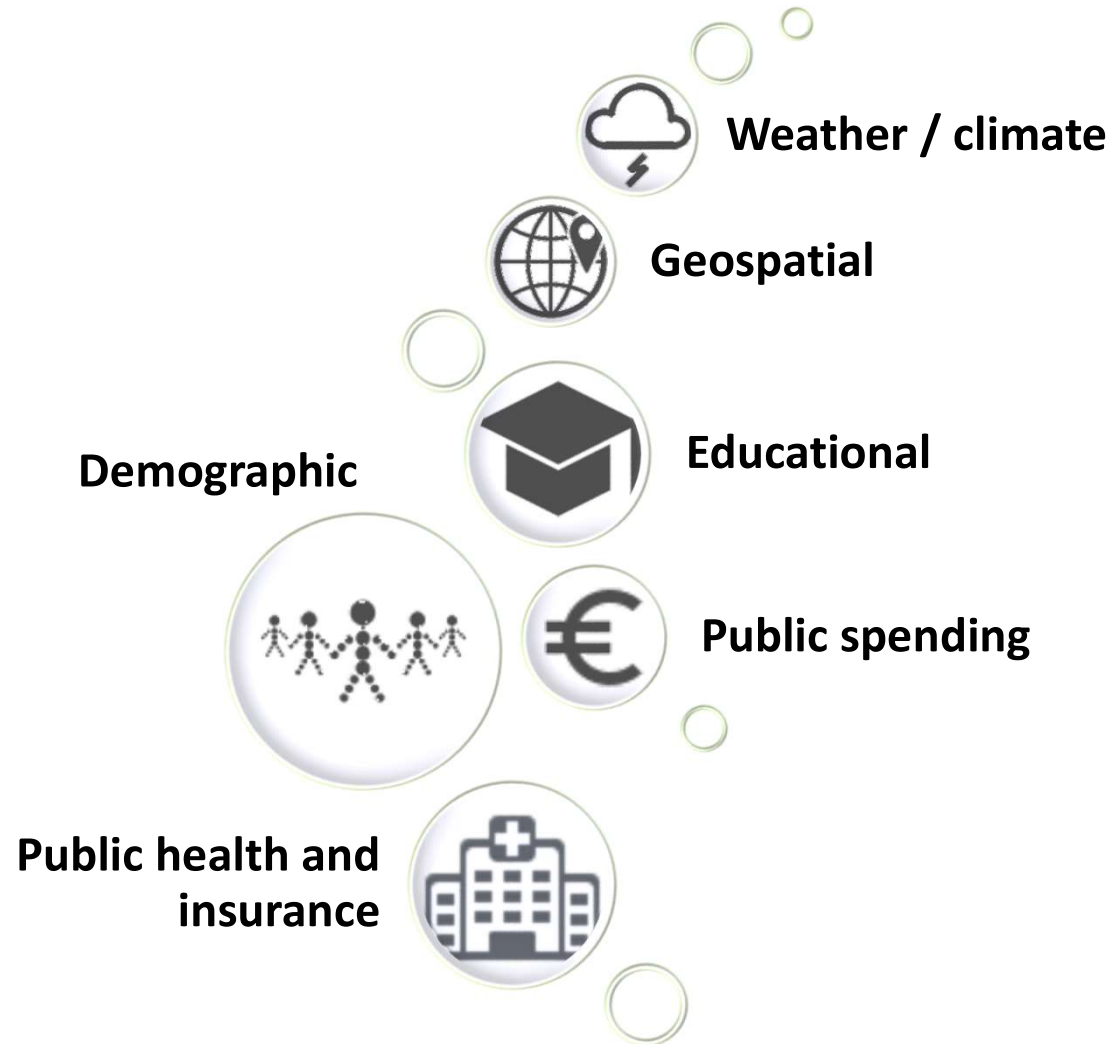
Maria Giagkou

Institute for Language and Speech Processing / Athena R.C.
ELRC



Connecting
Europe
Facility

The notion of data

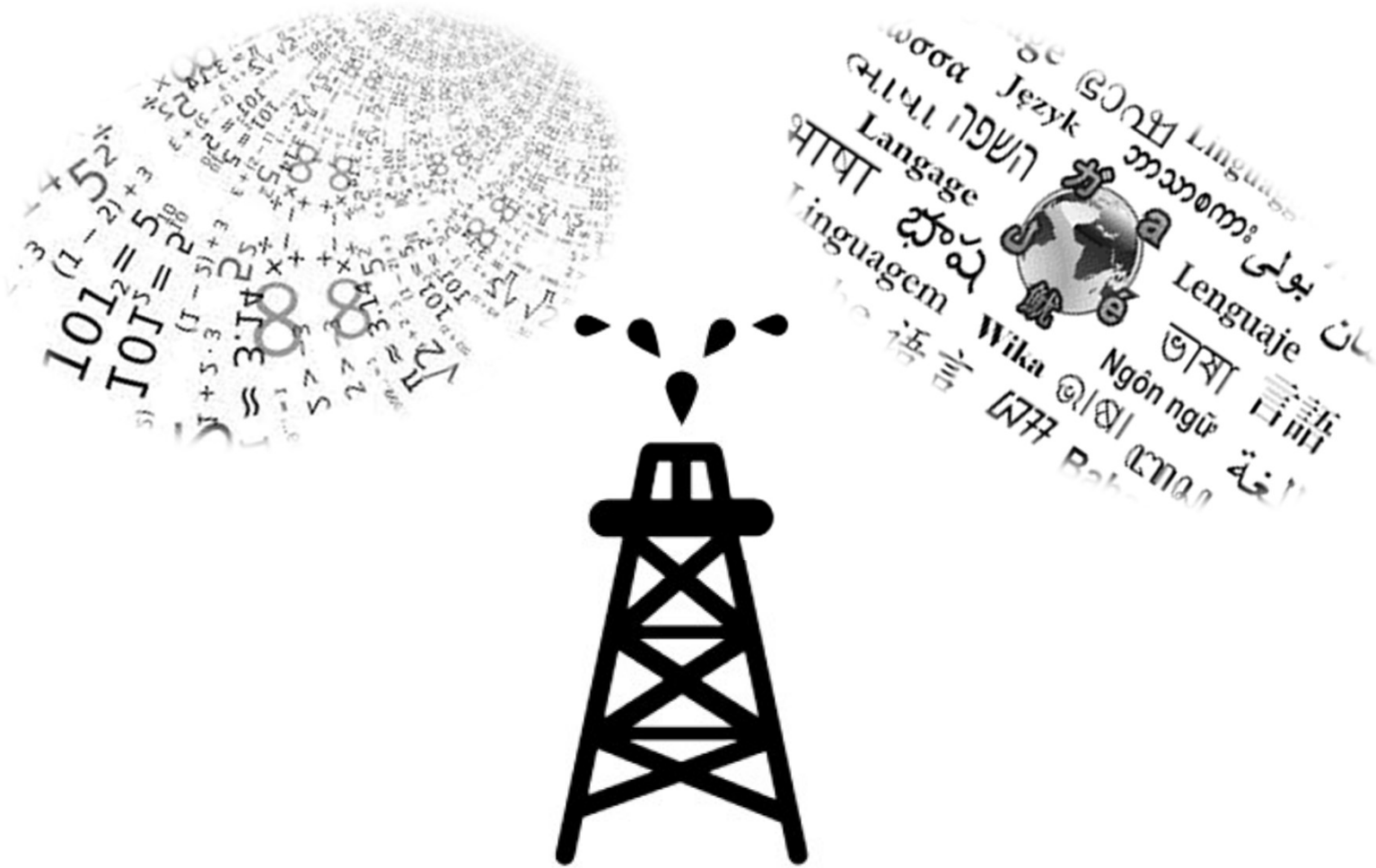


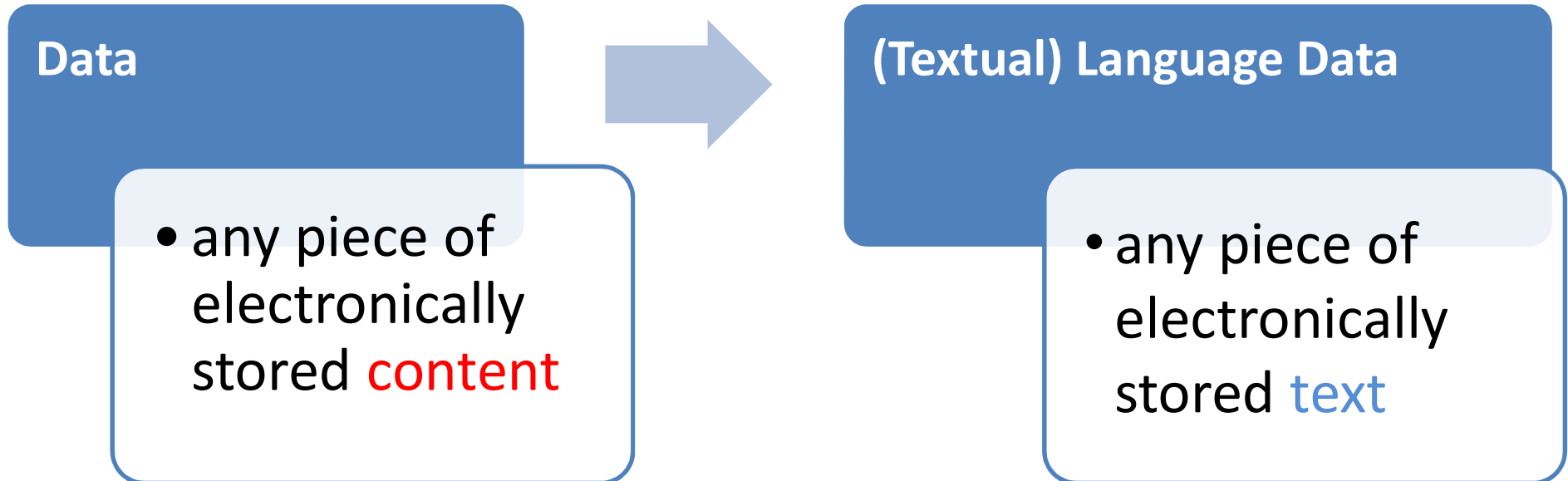


Data: the oil of the 21st century



The notion of data





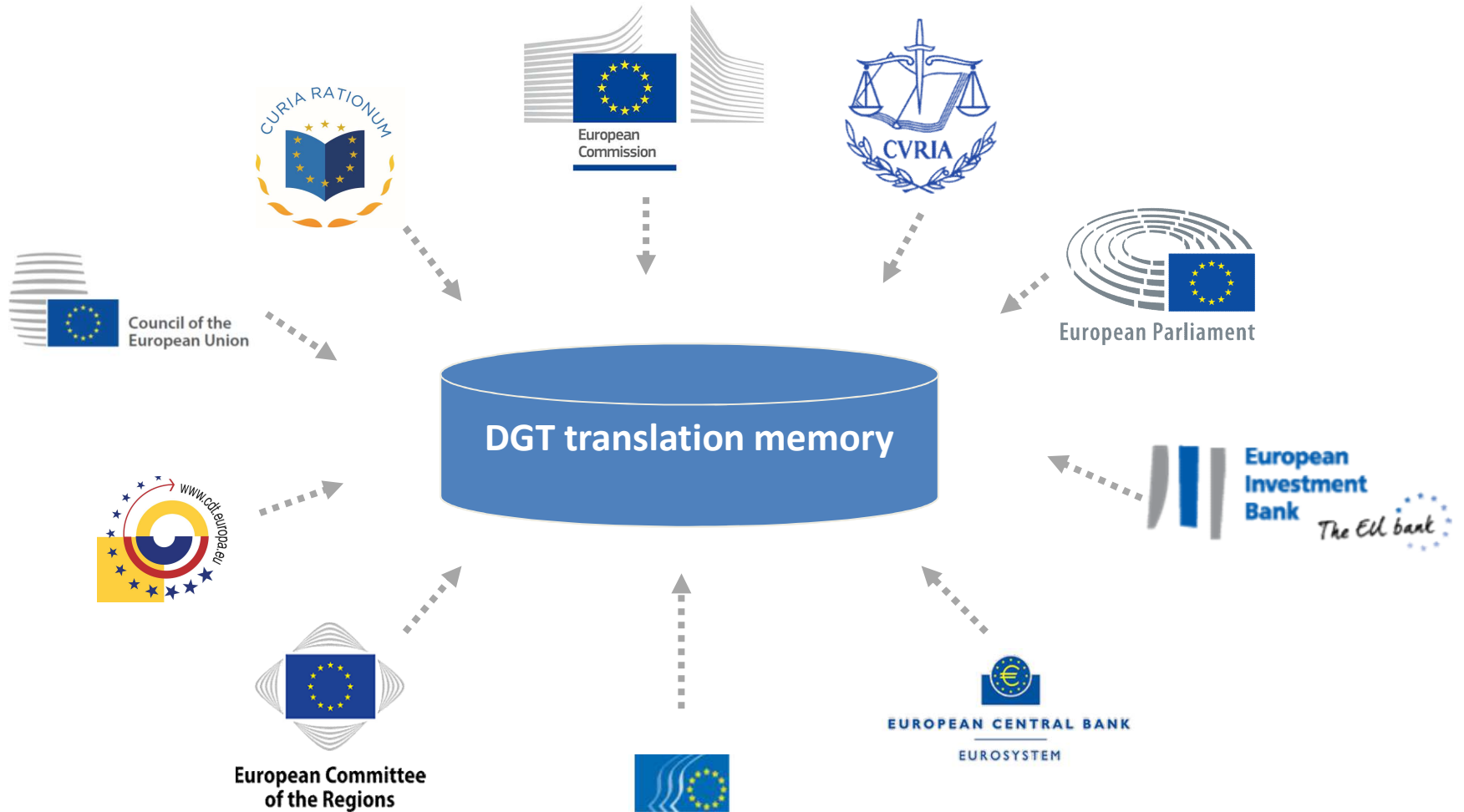
HR

Očuvana priroda doprinosi osiguravanju svih funkcionalnosti nužnih za život i ekonomski razvoj. U Republici Hrvatskoj, kao i u svijetu, priroda je pod stalnim pritiskom ljudskih djelatnosti. Iako se ulažu značajniji naponi za očuvanje prirode, pojedine njezine sastavnice su i dalje ugrožene.

EN

Conserved nature contributes to ensuring all functionalities necessary for livelihoods and economic development. In the Republic of Croatia, as well as in the world, nature is facing permanent pressures from human activities. Even though significant efforts are being invested in nature conservation, certain nature components are still being threatened.

Data used by eTranslation



Such data are already available
BUT
they are not enough...

What data are useful for eTranslation as per type | 1



- Any **electronically stored text** in an EU language plus NO and IS
- **Texts and their translations** (i.e. parallel bilingual or multilingual)

Croatian text

U ovome Statistikom izvješću Državni zavod za statistiku korisnicima stavlja na raspolaganje konane rezultate Popisa stanovništva, kuanstava i stanova u Republici Hrvatskoj 2011. godine. Ovo statistiko izvješe prikazuje podatke prema starosti i spolu po županijama, gradovima, opinama, naseljima te gradskim etvrtima Grada Zagreba. Podaci su prikazani sa stanjem na dan 31. ožujka 2011. u 24 sata (kritini trenutak Popisa) prema tada važeem teritorijalnom ustroju Republike Hrvatske. Popis stanovništva, kuanstava i stanova u Republici Hrvatskoj 2011. godine proveden je u razdoblju od 1. do 28. travnja 2011., prema stanju na dan 31. ožujka 2011. u 24 sata, što se smatra kritinim trenutkom Popisa.

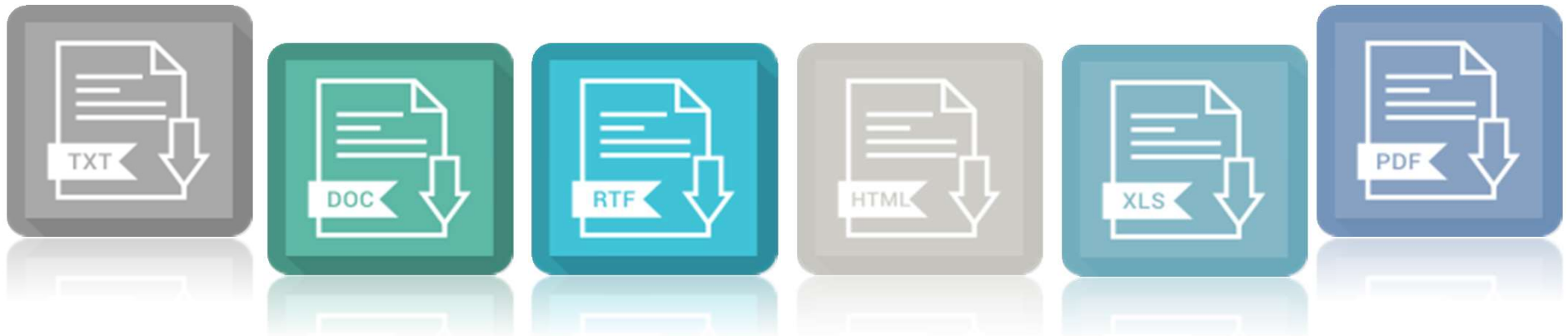
Translation in English

In this Statistical Report, the Croatian Bureau of Statistics puts at users' disposal the final data of the Census of Population, Households and Dwellings 2011. This Statistical Report presents data on age and sex by counties, towns, municipalities, settlements and the districts of City of Zagreb. Data are presented according to the situation as on 31 March 2011 at midnight (census moment) by the then relevant territorial constitution. The Census of Population, Households and Dwellings 2011 was carried out in the Republic of Croatia from 1 to 28 April 2011 according to the situation as on 31 March 2011 at midnight, which is deemed to be the census moment.

- List of terms and their translations, i.e. a **terminology**, e.g.

Croatian	English
aerodinamika	aerodynamics
zračna struja	air flow
zračna struja	air stream
granični sloj	boundary layer
čeonni otpor	drag
koeficijent otpora	drag coefficient
dinamički tlak	dynamic pressure
slobodna struja zraka	free air stream
aerodinamički uzgon	aerodynamic lift
koeficijent uzgona	lift coefficient
...	...

What data are useful for eTranslation as per format | 1



- In principle, any text in machine readable format
- But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing
- More processing introduces more errors in the final output, making it less useful for eTranslation

File formats for parallel texts

JANUARY - JUNE 2003

33

Minority Rights in Croatia

Dario Kuntić*

Original paper
UDC 323.15(497.5)
Received in May 2003

The aim of this paper is to present and to explain the level of protection of minority rights in the Republic of Croatia. As Croatia inherited from the former Socialist Federative Republic of Yugoslavia the system of the protection of rights of the minorities immediately after gaining independence, it provided a model of the realisation of minority rights. That model guaranteed minorities the right for education in their mother tongue, the right to the official use of their language and alphabet, the right to publishing in their mother tongue, access to the media, protection of cultural treasures and their entire cultural heritage, various venues of preservation of ethnic, language and religious identity and the representation of minority interests.

Key words: Republic of Croatia, minorities, minority rights, democratic principles

1. Introduction

By its Constitution, the Republic of Croatia bounds itself to protection of the minorities from any discrimination. They have the right to identity and participation in public affairs. Members of ethnic and national communities or minorities whose share in the total population is higher than 8% have the right for participation, proportional with their share in the entire population, in the Croatian Parliament, the Government of the Republic of Croatia and in the bodies of supreme judicial authorities.

Members of national minorities exercise their political rights, like all citizens or nationals in the Republic of Croatia through regular institutions on the basis of the provisions of the Constitution, Constitutional Law, Law on Election of Representatives to the Croatian Parliament, Law on Election of Members of Representation Bodies of Units of Self-government and Administration and Law on Political Parties.¹

The Republic of Croatia signed a number of international, bilateral and multilateral, contracts and

agreements related to the protection of human and minority rights such as the European Convention on Human Rights, the European Charter of Local Self-government, Charter on Regional and Minority Languages, the Framework Convention for the Protection of National Minorities, the Treaty on the Protection of the Italian Minority in the Republic of Croatia and the Croatian Minority in the Republic of Italy, and many others.

At the end of this introduction it would be important to mention that according to the 1991 census in Croatia there were 12,032 Albanians, 214 Austrians, 43,469 Muslims, 458 Bulgarians, 9,724 Montenegrins, 13,068 Czechs, 22,355 Hungarians, 6,280 Macedonians, 2,635 Germans, 679 Poles, 6,695 Roma, 810 Romanians, 706 Russians, 5,606 Slovaks, 581,663 Serbs, 21,303 Italians, 320 Turks, 2,494 Ukrainians, 22 Vlachs, 600 Jews and 3,012 other ethnic and national minorities.²

For the sake of comparison, according to the last census in 2001 in Croatia there were 15,082 Albanians, 247 Austrians, 20, 755 Muslims, 331 Bulgarians, 4,926 Montenegrins, 10,510 Czechs, 16, 595 Hungarians, 4,270 Macedonians, 2,902 Germans, 576 Poles, 9,463 Roma, 475 Romanians, 906 Rus-

* Dario Kuntić is a political scientist working in the Ministry of European Integration of the Republic of Croatia and holds an M.A. in Democratization and Human Rights from the University of Bologna, Italy



- The following formats are particularly useful (in descending order):
 - For bilingual/multilingual parallel texts
 1. Translation memories (.tmx)
 2. XML translation files (.xliff)
 3. Plain text (.txt, .csv)
 4. Spreadsheets (e.g. xlsx)
 - For terminologies
 1. TermBase eXchange (.tbx)
 2. Plain text (.txt, .csv)
 3. Spreadsheets (e.g. xlsx)
 - For monolingual texts
 1. Plain text (.txt, .csv)

File formats of parallel texts and their manipulation



Don'ts



This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.

Ovo je prethodni odlomak na hrvatskom jeziku. Ovo je prethodni odlomak na hrvatskom jeziku. Ovo je prethodni odlomak na hrvatskom jeziku. Ovo je prethodni odlomak na hrvatskom jeziku. Ovo je prethodni odlomak na hrvatskom jeziku. Ovo je prethodni odlomak na hrvatskom jeziku.

A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.

Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku.



Don'ts



English	Hrvatski
<p>This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.</p>	<p>Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku.</p>
<p>A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.</p>	<p>Drugi stavak lijevo na hrvatskom jeziku. Drugi stavak lijevo na hrvatskom jeziku. Drugi stavak lijevo na hrvatskom jeziku. Drugi stavak lijevo na hrvatskom jeziku. Drugi stavak lijevo na hrvatskom jeziku. Drugi stavak lijevo na hrvatskom jeziku.</p>





- filename01_EN.txt
- filename01_HR.txt
- filename02_EN.txt
- filename02_HR.txt
- filename03_EN.txt
- filename03_HR.txt
- filename04_EN.txt
- filename04_HR.txt
- filename05_EN.txt
- filename05_HR.txt
- filename06_EN.txt
- filename06_HR.txt
- filename07_EN.txt
- filename07_HR.txt
- filename08_EN.txt
- filename08_HR.txt
- filename09_EN.txt
- filename09_HR.txt
- filename10_EN.txt
- filename10_HR.txt

Use **identical filenames** for each document pair (source – translation)



Do's

- filename01_EN.txt
- filename01_HR.txt
- filename02_EN.txt
- filename02_HR.txt
- filename03_EN.txt
- filename03_HR.txt
- filename04_EN.txt
- filename04_HR.txt
- filename05_EN.txt
- filename05_HR.txt
- filename06_EN.txt
- filename06_HR.txt
- filename07_EN.txt
- filename07_HR.txt
- filename08_EN.txt
- filename08_HR.txt
- filename09_EN.txt
- filename09_HR.txt
- filename10_EN.txt
- filename10_HR.txt

Include **language identifiers** in the filename



- A dataset is a collection of data **grouped according to certain criteria**
- For the purpose of enhancing and adapting CEF eTranslation, two criteria are critical:
 - **Language(s)**: each collection is defined by the language or language pairs of its data, e.g.
 - *Collection of texts in English – Croatian*
 - *Documents in English – Croatian - French*
 - **Domain**: each collection ideally belongs to a single domain, e.g.
 - *Collection of texts in English – Croatian in the culture domain*
 - *Social security documents in English – Croatian - French*



- Administrative/regulatory domain and
- Topics relevant to the CEF DSIs

CEF DSI	Domain
Online Dispute Resolution	Consumers' rights, complaints
Electronic Exchange of Social Security Information	Social security, insurance
eProcurement	Public procurement, contractual agreements
European e-Justice Portal	Justice, Law
eHealth	Health, Medicine
Business Registers Interconnection System	Business, market
Safer Internet	
Cybersecurity	
Public Open Data	
Europeana	Culture

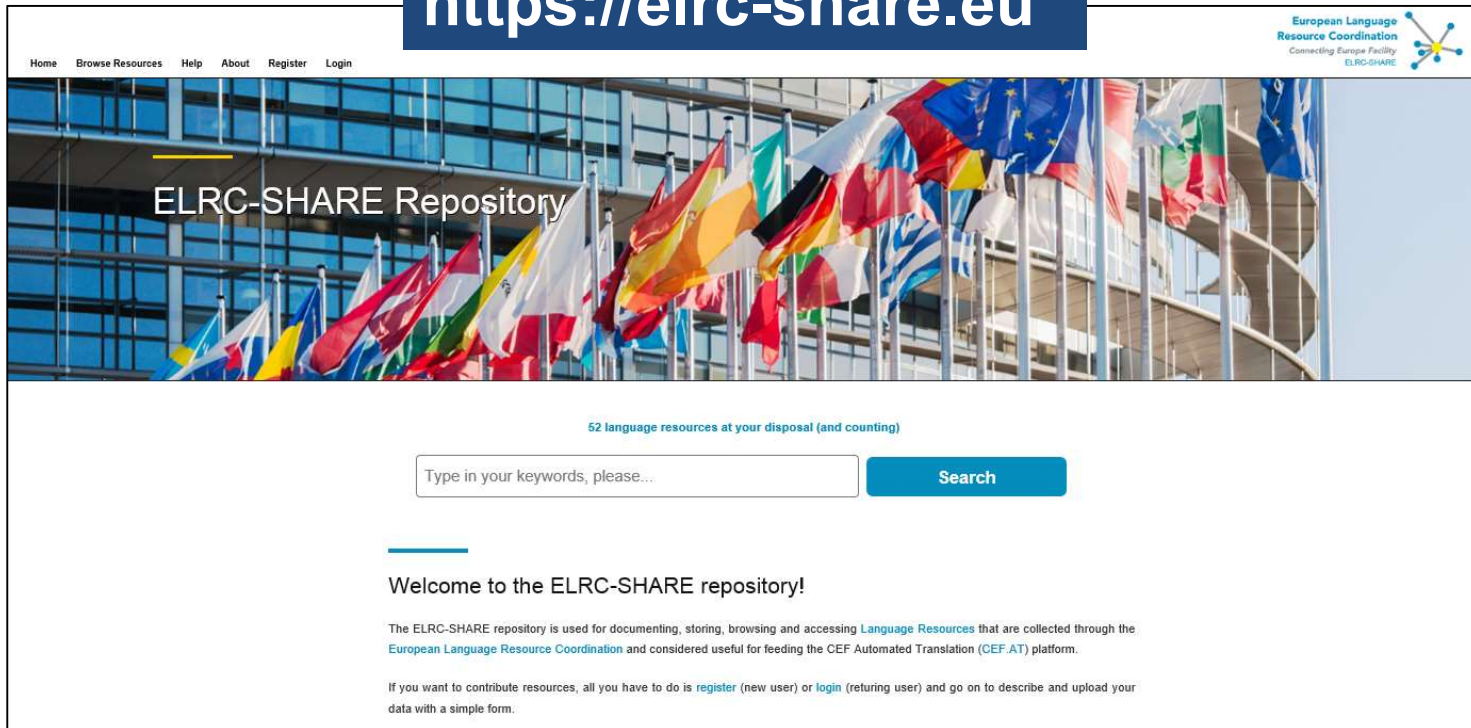
How to contribute your data to CEF eTranslation

A step-by-step guide

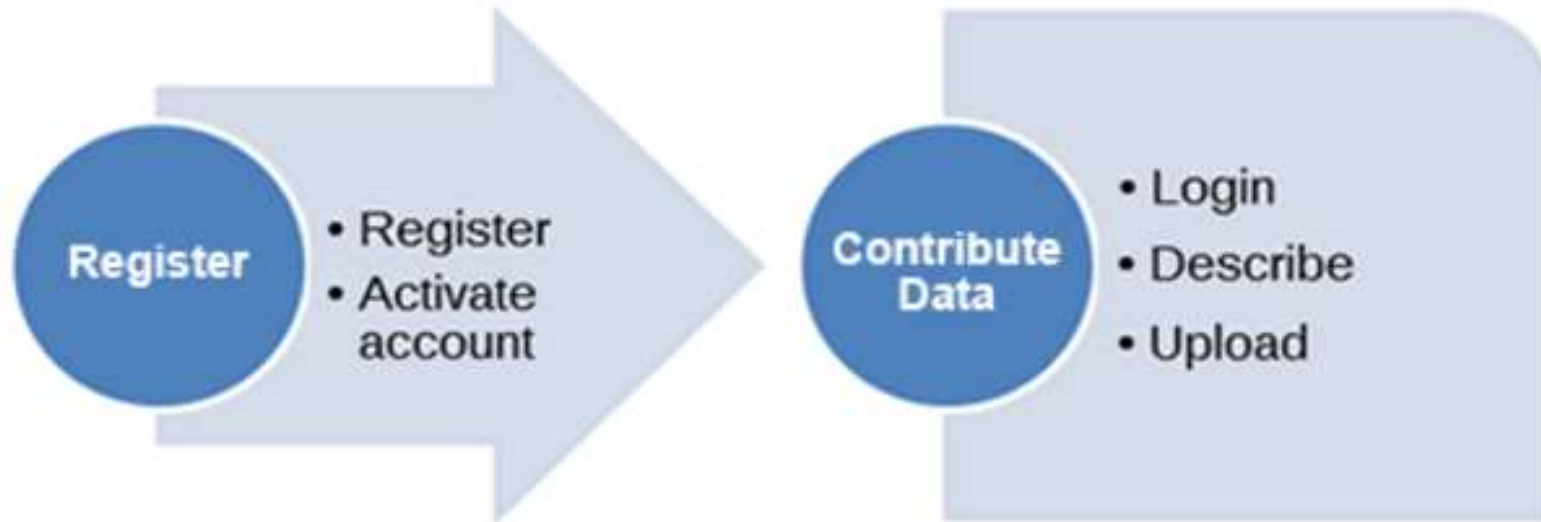
The ELRC-SHARE Repository

- Access to, sharing and contribution of language resources
- Access to tools and services catalogue

<https://elrc-share.eu>



The screenshot shows the homepage of the ELRC-SHARE Repository. At the top, there is a navigation menu with links for Home, Browse Resources, Help, About, Register, and Login. The main header features the text "ELRC-SHARE Repository" overlaid on a background image of various European national flags. Below the header, a blue box indicates "52 language resources at your disposal (and counting)". A search bar with the placeholder text "Type in your keywords, please..." and a blue "Search" button is provided. The main content area includes a welcome message: "Welcome to the ELRC-SHARE repository!" followed by a paragraph explaining the repository's purpose: "The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF Automated Translation (CEF.AT) platform." A final paragraph states: "If you want to contribute resources, all you have to do is register (new user) or login (returning user) and go on to describe and upload your data with a simple form."



How to Register (1/2)



 Register

[Home](#) [Browse Resources](#) [Help](#) [About](#) [Register](#) [Login](#)

ELRC-SHARE Repository

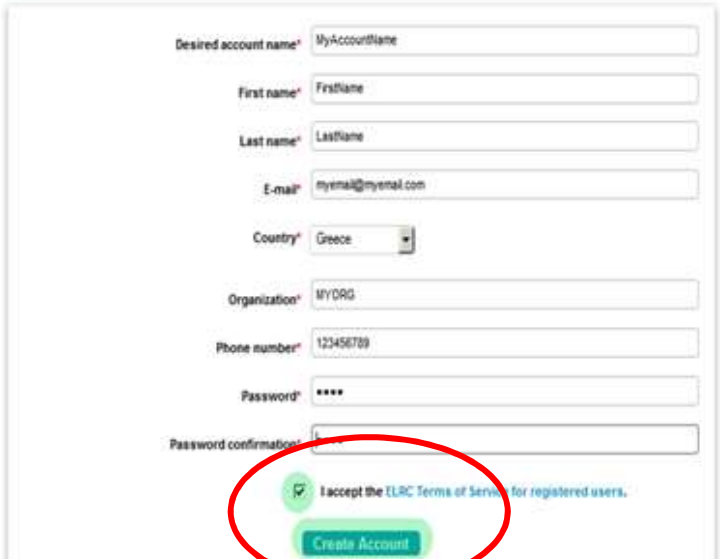


Welcome to the ELRC-SHARE repository!



- Fill in the required info
- Read the *Terms of Service* and click *Accept*, if you agree
- Click the *Create Account* button
- Activate your account according to the guidelines emailed to you

*All fields are required



Desired account name* MyAccountName

First name* FirstName

Last name* LastName

E-mail* myemail@myemail.com

Country* Greece

Organization* MYORG

Phone number* 123456789

Password* ****

Password confirmation*

I accept the ELRC Terms of Service for registered users.

Create Account




Data Contribution

New Resource

Resource Title*

The name by which the resource is already known or by which you would like it to be known; e.g. "The GSRT bilingual corpus of Greek-English bulletins"

- Fill in the details of the dataset



The screenshot shows a web form with three main sections:

- Resource Title***: A text input field containing "Bilingual resource name". Below it is a descriptive paragraph: "The name by which the resource is already known or by which you would like it to be known; e.g. 'The GSRT bilingual corpus of Greek-English bulletins'".
- Resource short description***: A text area containing "A short resource description:". Below it is a descriptive paragraph: "A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, news bulletins, lexicon of sports terminology etc.)".
- Language(s)**: A dropdown menu with a scroll bar. The visible options are: Croatian, Danish, Dutch, Flemish, English (highlighted in blue), Estonian, Finnish, French (highlighted in blue), German, and Hungarian.

- Three modes for contributing your data

Contribution Mode*

- Upload ZIP archive
- Provide URL of resources
- eDelivery (Generate XML file to attach to your eDelivery contribution)

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

Upload Resource*

Choose File No file chosen

Please upload a **.zip file** up to 100MB.

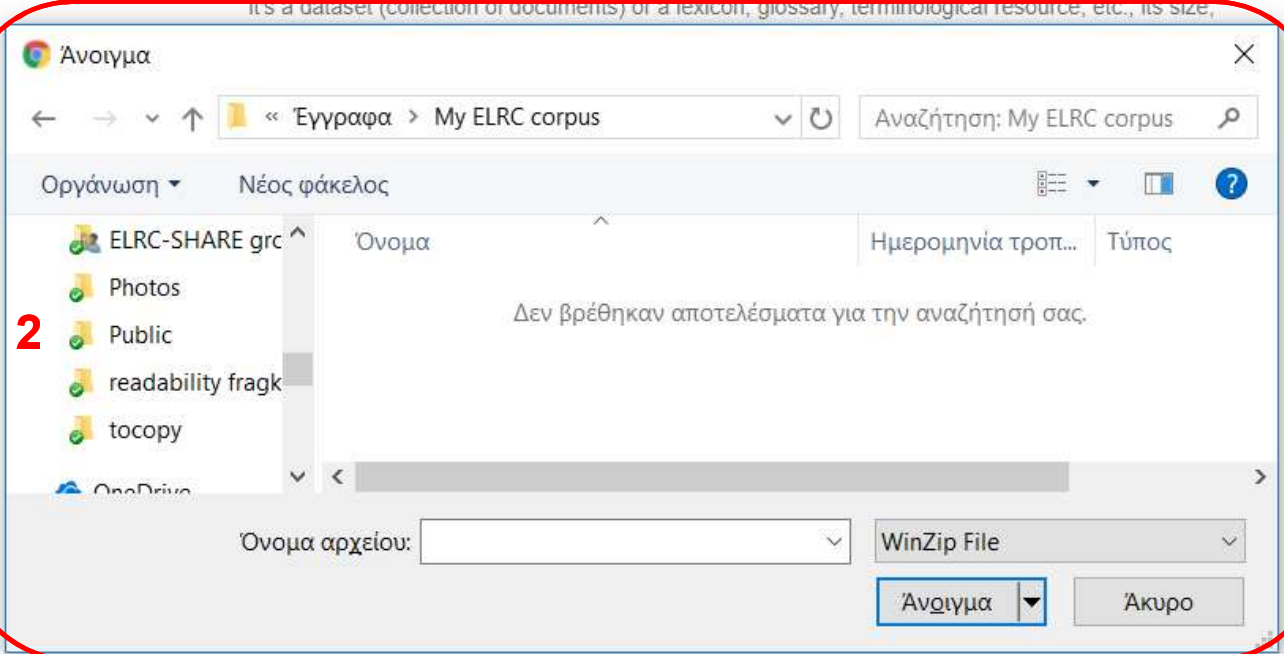
In case the **.zip file** file you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr

Submit

Reset

1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit

A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size,



2

1 Choose File No file chosen
Please upload a .zip file up to 100MB.
In case the .zip file you wish to upload is larger than 100MB, please contact elrc-share@lsp.gr

3 Submit Reset



- Alternatively indicate a url (directory listing)

Language(s)*

Bulgarian
Czech
Croatian
Danish
Dutch; Flemish
English
Estonian
Finnish
French
German
Hungarian

The language(s) of the resource; for resources with multiple languages, hold down CTRL key to select multiple values

Contribution Mode*

Upload ZIP archive
 Provide URL of resources

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

Resource URL*

www

Please provide a URL containing the files you wish to contribute



Contribute Your Data Through eDelivery

If you wish to share your data through [eDelivery](#), you can use the ELRC-SHARE CEF eDelivery Access Point. For more information click [here](#). In such a case, please fill in the form on the left and choose eDelivery in the Contribution mode.



Help

Documentation on the ELRC-SHARE editor

The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LRs:

- [Walkthrough for contributors](#)
- [Walkthrough for editors](#)

ELRC-SHARE schema

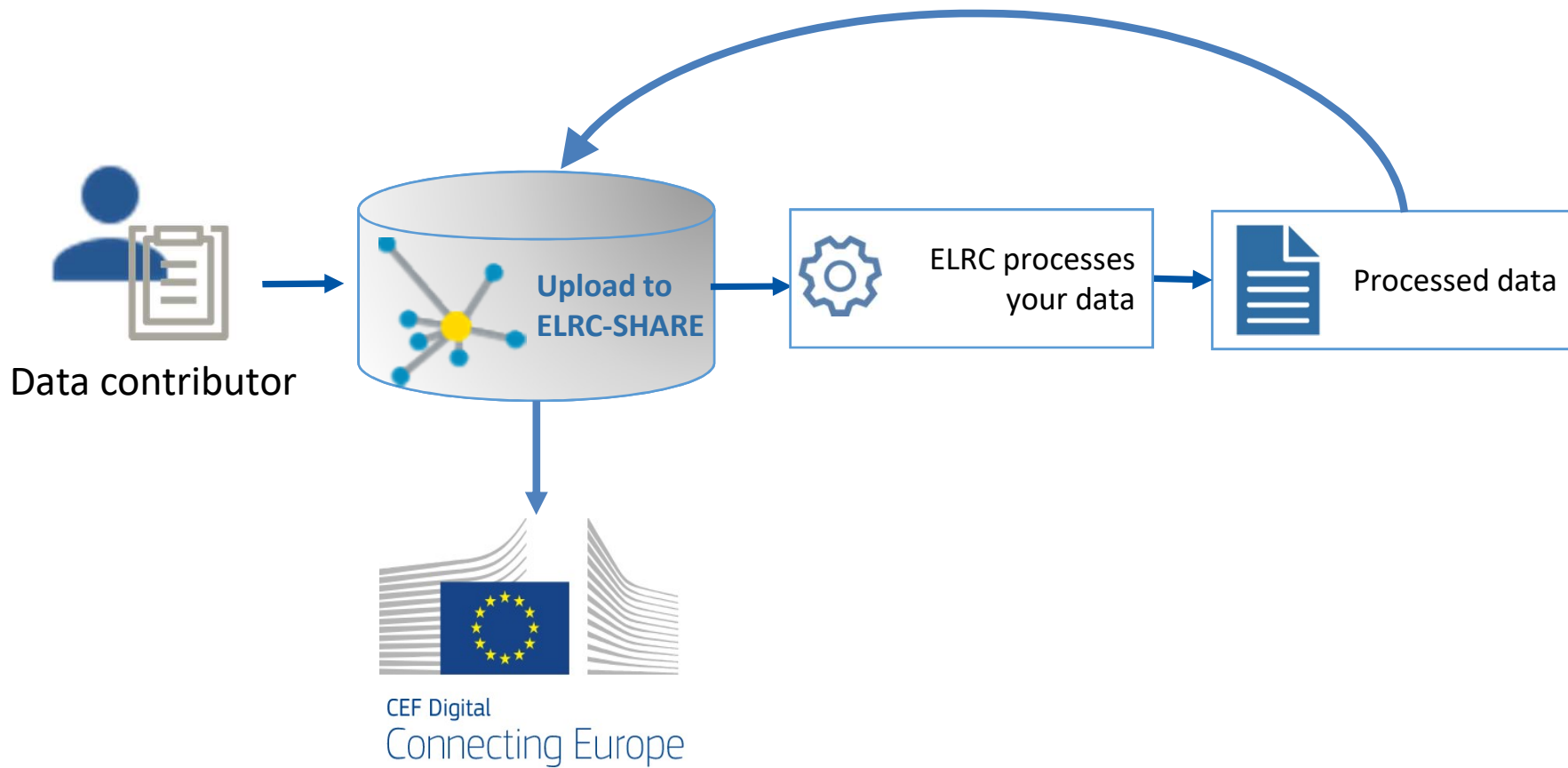
- [ELRC-SHARE schema XSD](#) (based on the META-SHARE Schema)
- [Documentation about the schema](#)

Preparing and sharing data with the
ELRC-SHARE repository

and what happens next



What happens to your data?





Data extraction

If your data is trapped in archives and databases, we can help extract it



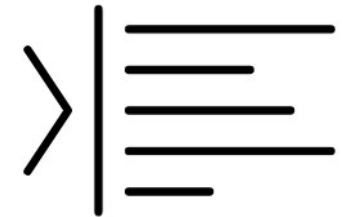
Anonymisation

Does your data contain private info? We can help to anonymise



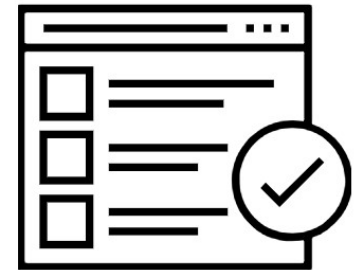
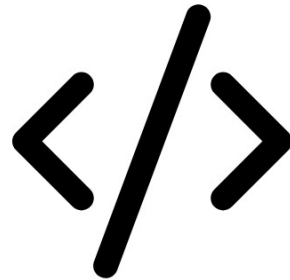
Cleaning

If your data is messy (i.e., lots of noise), we will clean it up



Re-formatting

Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!



Data conversion

If your data isn't converted to the proper formats, we can help convert it

Tag removal

Does your data contain unneeded tags? We can assist in removing them!

Alignment

Translations aren't aligned? We'll do it for you with our tools!

Metadata

Metadata are crucial! We can organise and validate metadata for your team

What has happened to your data?

File01_hr.txt
File01_en.doc
File02_hr.pdf
File02_en.txt
File03_hr.doc
File03_en.doc
...

After
processing

```
<tu tuid="818">
  <tuv xml:lang="hr">
    <seg>U Republici Hrvatskoj, kao i u svijetu,
    priroda je pod stalnim pritiskom ljudskih
    djelatnosti.</seg> </tuv>
  <tuv xml:lang="en">
    <seg>In the Republic of Croatia, as well as in
    the world, nature is facing permanent pressures from
    human activities.</seg></tuv></tu>
  <tu tuid="819">
    <tuv xml:lang="hr">
      <seg>Iako se ulažu značajniji naponi za očuvanje
      prirode, pojedine njezine sastavnice su i dalje
      ugrožene.</seg></tuv>
    <tuv xml:lang="en">
      <seg>Even though significant efforts are being
      invested in nature conservation, certain nature
      components are still being threatened.</seg></tuv></tu>
```

How your dataset is described



Bilingual Croatian-English Parallel Corpus (Processed)

Bilingual Croatian-English Parallel Corpus of 21340 translation units in the public administration domain.

[← Back](#) [Download](#) [Edit Resource](#)

Distribution

Availability: Available

Licences

Terms for PSI-compliant resources

Open Under-PSI

Distribution Details

IPR Holders

- Ministry of the Interior of the Republic of Croatia
- Ministry of Environment and Energy of the Republic of Croatia
- Government of the Republic of Croatia
- Ministry of the Sea, Transport and Infrastructure of the Republic of Croatia
- Ministry of Defence of the Republic of Croatia of the Republic of Croatia
- Ministry of Culture of the Republic of Croatia of the Republic of Croatia
- Ministry of Construction and Physical Planning of the Republic of Croatia
- Central State Office for Croats Abroad
- Ministry of Finance of the Republic of Croatia
- Institute of Public Finance of the Republic of Croatia

Contact Person

Marko Tadić

text

Bilingual text corpus

Languages

Croatian (hr) (376,779 Words)

English (en) (439,136 Words)

Linguality

Linguality type: Bilingual

Multi-linguality type: Parallel

Text Format

TMX

Size

21,340 Translation Units

Character encoding

UTF-8

Domains

POLITICS

Executive Power And Public Service (Eurovoc 0436)

Annotation

Alignment

StandOff: False

Segmentation level: Sentence

Standard practices conformance: TMX

Annotation Mode: Automatic

Annotation Tools:

- ILSP-FC alignment and TMX filtering module

Resource Creation

Resource Creator

Marko Tadić

Creation lasted: 01/02/2018 - 01/02/2018

Funding Project

European Language Resource Coordination LOT3 (ELRC Data - Tools and Resources for CEF Automated Translation - LOT3 (SMART 2015/1091 - 30-CE-0816766/00-92))

URL: <http://www.lr-coordi...>

Funding Type: Service Contract

Funder: European Commission

Funding Country: European Union (EU)

Project duration: 13/12/2016 - 12/02/2020

Metadata

Created: 25/03/2017

Metadata Language: English (en)

Version

Version: 2.0

Last Updated: 01/02/2018

Relations

Related Resource: Bilingual Croatian-English Parallel Corpus

Relation Type: Is Version Of



Bilingual Croatian-English Parallel Corpus (Processed)

Bilingual Croatian-English Parallel Corpus of 21340 translation units in the public administration domain.

[← Back](#) [Download](#) [Edit Resource](#)

Distribution

Availability: Available

Licences

Terms for PSI-compliant resources

Open Under-PSI

Distribution Details

IPR Holders

- Ministry of the Interior of the Republic of Croatia
- Ministry of Environment and Energy of the Republic of Croatia
- Government of the Republic of Croatia
- Ministry of the Sea, Transport and Infrastructure of the Republic of Croatia
- Ministry of Defence of the Republic of Croatia of the Republic of Croatia
- Ministry of Culture of the Republic of Croatia of the Republic of Croatia
- Ministry of Construction and Physical Planning of the Republic of Croatia
- Central State Office for Croats Abroad
- Ministry of Finance of the Republic of Croatia
- Institute of Public Finance of the Republic of Croatia

Contact Person

Marko Tadić

text

Bilingual text corpus

Languages

Croatian (hr) (376,779 Words)

English (en) (439,136 Words)

Linguality

Linguality type: Bilingual

Multi-linguality type: Parallel

Text Format

TMX

Size

21,340 Translation Units

Character encoding

UTF-8

Domains

POLITICS

Executive Power And Public Service (Eurovoc 0436)

Annotation

Alignment

StandOff: False

Segmentation level: Sentence

Standard practices conformance: TMX

Annotation Mode: Automatic

Annotation Tools:

- ILSP-FC alignment and TMX filtering module

Resource Creation

Resource Creator

Marko Tadić

Creation lasted: 01/02/2018 - 01/02/2018

Funding Project

European Language Resource Coordination LOT3 (ELRC Data - Tools and Resources for CEF Automated Translation - LOT3 (SMART 2015/1091 - 30-CE-0816766/00-92))

URL: <http://www.lr-coordi...>

Funding Type: Service Contract

Funder: European Commission

Funding Country: European Union (EU)

Project duration: 13/12/2016 - 12/02/2020

Metadata

Created: 25/03/2017

Metadata Language: English (en)

Version

Version: 2.0

Last Updated: 01/02/2018

Relations

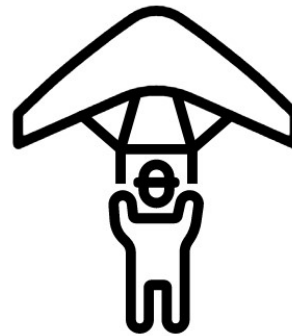
Related Resource: Bilingual Croatian-English Parallel Corpus

Relation Type: Is Version Of



All these services can also be offered on-site to all data contributors free of charge



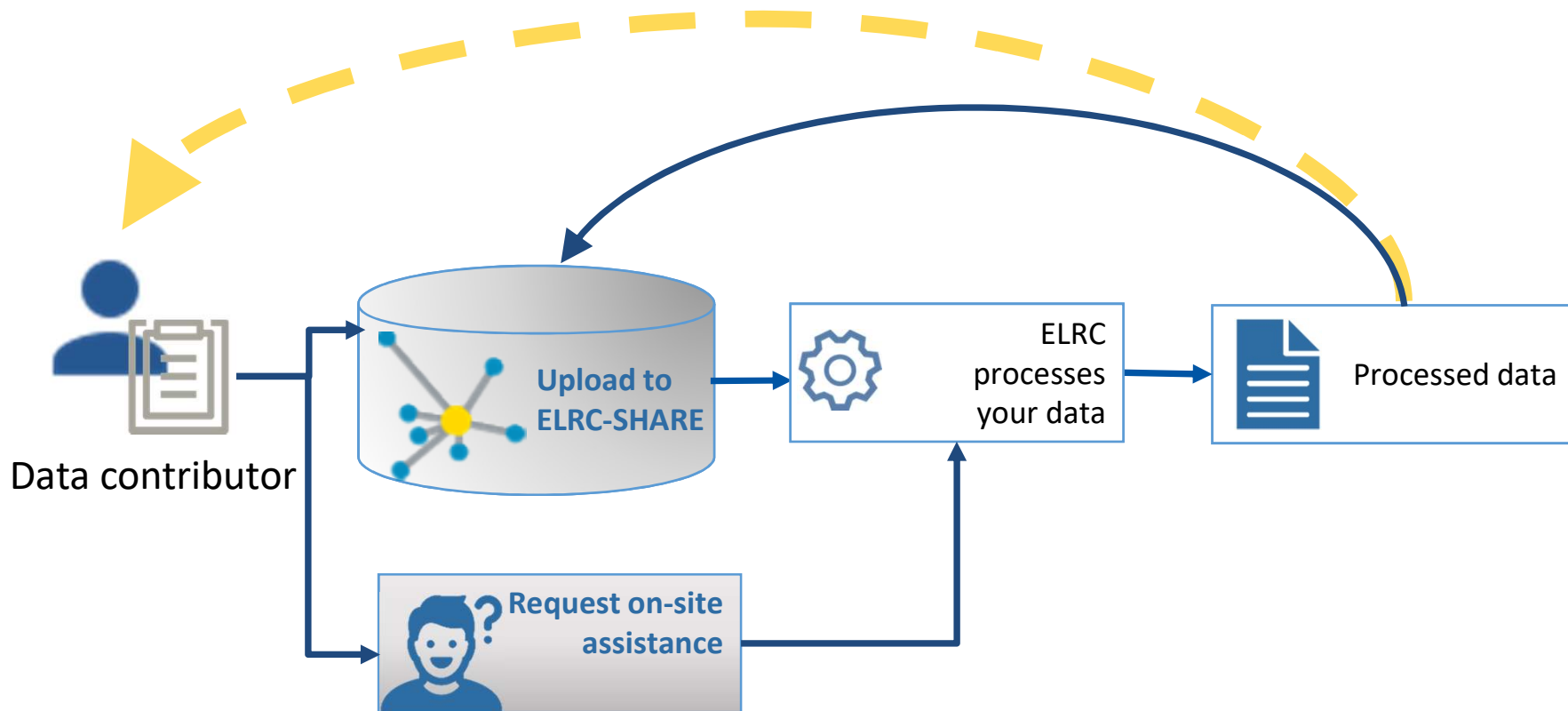


**Our team of experts will travel
directly to assist you
at your own offices**



We will fix your data issues and return the processed data directly to you. We can also help to improve your data management processes. Just ask!

What happens to your data?



How to request services and help



www.lr-coordination.eu/request-onsite-assistance

Submit a request for on-site assistance by filling out the form below. See a list of services [here](#).

First name *

Last name *

Institution *

Country *

Email *

Types of assistance required *

- Legal assistance
- Data processing
- Anonymisation
- Other

Description of assistance required

Submit

www.lr-coordination.eu/helpdesk

[Home](#) [Discover](#) [Resources](#) [Services](#) [Events](#) [Anchor Points](#) [News](#) [Helpdesk](#)



Please feel free to contact us through one of the following channels:

Telephone* **+33 970 440 522**

Secretariat Support **+49 681 857 7552 85**

Skype **ELRC Helpdesk**

E-mail help@lr-coordination.eu

Hvala vam!



- By [Michael Mellon](#), GB, , CC-BY 3.0 US
- By [Joana Pereira](#), BR, CC-BY 3.0 US
- By [Becca O'Shea](#), NZ, CC-BY 3.0 US
- By [Creative Stall](#), Basic licence www.iconfinder.com
- By [Creative Stall](#), PK, CC-BY 3.0 US
- By [Arthur Shlain](#), IL, CC-BY 3.0 US
- By [Shmidt Sergey](#), US, CC-BY 3.0 US
- By [Gregor Cresnar](#), CC-BY 3.0 US
- By [anbileru adaleru](#), CC-BY 3.0 US
- By [Vectors Market](#), CC-BY 3.0 US